

BAB I

PENDAHULUAN

1.1 Latar Belakang

Kebutuhan akan informasi yang cepat dan akurat telah menempatkan media menjadi salah satu kebutuhan penting dalam masyarakat. Semakin menjamurnya media informasi saat ini mengakibatkan jumlah informasi terus meningkat. Keadaan ini disebut *Information overload* yaitu istilah yang menggambarkan suatu keadaan dimana efisiensi seseorang saat menggunakan informasi dalam pekerjaannya terhambat oleh jumlah informasi yang relevan dan berpotensi berguna (Bawden & Robinson, 2008).

Seiring meningkatnya kebutuhan akan informasi, maka berbanding lurus dengan jumlah dokumen yang ada di dunia internet, salah satunya adalah dokumen berita. Berita adalah kumpulan informasi tentang banyak peristiwa penting yang terjadi dan diperbaharui secara berkala. Memahami isi dokumen berita melalui ringkasan teks memerlukan waktu yang lebih singkat dibandingkan membaca seluruh isi dokumen, sehingga ringkasan teks menjadi sangat penting. Dengan adanya ringkasan, diharapkan pembaca dapat dengan cepat dan mudah dalam memahami pokok pikiran dari sebuah teks tanpa harus membaca keseluruhan teks. Selain dapat menghemat waktu, pembaca juga dapat menghindari pembacaan teks yang tidak relevan dengan informasi yang diharapkan oleh pembaca, terutama ketika ada banyak sekali informasi yang tersedia di internet. Oleh karena itu, maka perlu menggunakan metode untuk

memahami, mengklasifikasikan dan menyajikan semua informasi dengan cara yang jelas dan ringkas, memungkinkan pembaca untuk menghemat waktu dan sumber daya (Ferreira, et al., 2013).

Ada dua teknik dalam peringkasan teks yaitu teknik ekstraksi dan abstraksi. Teknik ekstraksi merupakan suatu teknik menyalin bagian teks yang paling penting dari teks sumber menjadi sebuah ringkasan, sedangkan abstraksi adalah teknik yang digunakan untuk mengambil intisari dari teks sumber kemudian membuat ringkasan dengan menciptakan kalimat-kalimat baru yang merepresentasikan intisari dari teks sumber dalam bentuk yang lain (Gupta & Gambhir, 2016). Penelitian awal dalam bidang peringkasan teks secara otomatis dimulai dengan pembuatan metode *term frequency* oleh Luhn pada tahun 1958. Metode Luhn didasarkan pada asumsi bahwa frekuensi kata di dalam sebuah dokumen teks merupakan indikasi terhadap tingkat kepentingannya (Jezek & Steinberger, 2008). Setelah itu, muncul berbagai metode diantaranya Latent Semantic Analysis (Yeh, Ke, Yang, & Meng, 2005), Machine Learning (Gupta & Gambhir, 2016), Genetic Algorithm (Fattah & Ren, 2009), Graph (Kumar & Salim, 2012), Sentence Scoring (Ferreira, et al., 2013), dan Naive-Bayes (Najibullah, 2015).

Dalam penelitian ini, bahasa yang digunakan dalam peringkasan teks adalah bahasa Indonesia. Indonesia adalah negara dengan populasi terbesar keempat di dunia. Dengan lebih dari 230 juta penutur, bahasa Indonesia menjadi salah satu bahasa yang paling banyak digunakan di dunia (Najibullah, 2015).

Penelitian dalam bidang peringkasan teks untuk bahasa Indonesia belum banyak dilakukan. Penelitian oleh Budhi dkk pada 2007 menerapkan metode Graph dan Algoritma Exhaustive (Budhi, Intan, Silvia, & Stevanus, 2007). Selanjutnya pada 2012, Aristoteles dkk melakukan penelitian dengan menerapkan metode Algoritma Genetika (Aristoteles, Herdiyeni, Ridha, & Adisantoso, 2012). Penelitian berikut pada 2014 oleh Silvia dkk menggunakan metode Latent Dirichlet Allocation dan Algoritma Genetika (Silvia, et al., 2014). Beberapa penelitian di atas sudah baik dilakukan, namun terdapat masalah yaitu rata-rata tingkat akurasi masih dibawah 55%.

Dalam peringkasan teks secara otomatis, pembobotan kalimat salah satu merupakan bagian penting. Pada 2013, Ferreira dkk melakukan evaluasi terhadap 15 fitur pembobotan kalimat untuk menentukan fitur mana yang lebih optimal. Hasil penelitian menunjukkan bahwa tiap fitur tersebut memiliki tingkat pengaruh yang berbeda-beda terhadap hasil ringkasan sistem. Selain itu, jumlah fitur yang banyak juga menyebabkan waktu komputasi lebih lama, sehingga perlu dipilih fitur-fitur yang dapat mempersingkat waktu sistem untuk menghitung bobot setiap kalimat (Ferreira, et al., 2013). Jika waktu penghitungan bobot tiap kalimat bisa ditekan menjadi lebih cepat, maka dapat berimplikasi pada total waktu peringkasan di setiap dokumen. Oleh karena itu, dalam penelitian ini akan digunakan 8 fitur pembobotan kalimat yang relevan dengan karakteristik dari sebuah teks berita.

Text processing adalah bidang penelitian tentang pengolahan data yang samar-samar. Ada banyak metode yang bisa dikembangkan untuk hal tersebut.

Sebagai contoh, *Rough clustering* adalah metode yang sangat baik untuk pemrosesan data yang samar. Namun, *supervised classification* lebih mudah daripada pendekatan *clustering* karena *clustering* yang dipartisi memerlukan beberapa metode awal untuk mendapatkan hasil yang lebih baik (Setyohadi, Abu Bakar, & Othman, 2010). *Decision tree* adalah salah satu metode klasifikasi yang menggunakan representasi struktur tree, dimana masing-masing node mewakili atribut yang diuji. Algoritma ID3 dan C4.5 telah diperkenalkan oleh J.R Quinlan yang menghasilkan pohon keputusan yang baik. Menurut (Hssina, Merbouha, Ezzikouri, & Erritali, 2014), algoritma C4.5 lebih baik dari sisi akurasi dan kecepatan proses komputasi jika dibanding dengan algoritma ID3.

Berdasarkan analisis penelitian terdahulu, maka dalam penelitian ini akan mengkombinasikan metode *sentence scoring* dan *decision tree* untuk peringkasan teks secara otomatis dalam bahasa Indonesia. Algoritma decision tree C4.5 digunakan untuk memilih kalimat mana saja yang akan masuk dalam ringkasan sistem. Untuk menghasilkan rule-rule dalam pohon keputusan, akan digunakan 50 teks berita sebagai data training. Model yang dihasilkan pada tahap training diterapkan pada proses seleksi kalimat yang akan masuk dalam ringkasan sistem. Ringkasan yang dihasilkan akan diuji tingkat akurasinya dengan mengukur nilai precision, recall, dan f-measure kemudian dibandingkan dengan penelitian sebelumnya.

1.2 Rumusan Masalah

Berdasarkan uraian diatas, maka dirumuskan permasalahan yang menjadi acuan dalam penelitian ini adalah sebagai berikut:

1. Pentingnya suatu sistem yang dapat meringkas teks secara otomatis untuk menjawab masalah efisiensi waktu baik dalam pembuatan ringkasan, maupun dalam memahami isi dari sebuah teks bacaan.
2. Perlu dilakukan peningkatan akurasi hasil ringkasan secara otomatis dengan metode yang diusulkan agar lebih baik dari metode sebelumnya yang diterapkan pada bahasa Indonesia.

1.3 Batasan Masalah

Untuk mencegah meluasnya cakupan permasalahan yang akan dibahas dalam penelitian ini, maka perlu dilakukan pembatasan masalah sebagai berikut :

1. Penelitian dalam lingkup peringkasan teks secara otomatis terdiri dari dua jenis masukan pada sistem yaitu *multi-document* dan *single-document*, maka dalam penelitian ini penulis membatasi jenis masukan berupa dokumen teks tunggal atau *single-document*.
2. Objek yang menjadi fokus dalam penelitian ini adalah teks berita bahasa Indonesia.
3. Data yang digunakan sebagai data training yaitu 50 teks berita nasional yang bersumber dari berita online harian Kompas yang merupakan korpus penelitian Aristoteles dkk pada tahun 2012.

1.4 Tujuan Penelitian

Penelitian ini pada dasarnya dilaksanakan untuk mencapai tujuan yaitu menghasilkan suatu sistem peringkasan teks bahasa indonesia secara otomatis dengan mengkombinasikan metode *sentence scoring* dan *decision tree*. Selain itu penelitian ini juga bertujuan untuk melakukan optimasi peringkasan teks, dimana hasil pengujian menggunakan metode yang diusulkan dalam penelitian ini akan dibandingkan dengan hasil pengujian dari metode lain yang diterapkan pada bahasa indonesia.

1.5 Manfaat Penelitian

Penelitian ini diharapkan akan memberikan manfaat sebagai berikut :

1. Menghasilkan ringkasan teks yang relevan dan melakukan peringkasan teks secara cepat dan menghemat waktu untuk memperoleh intisari dokumen.
2. Memberikan efisiensi waktu bagi para pembaca dalam memperoleh informasi melalui hasil ringkasan dokumen.
3. Dapat bermanfaat untuk para pengembang sistem yang ingin membuat aplikasi peringkasan teks secara otomatis dengan mengadopsi metode yang diusulkan dalam penelitian ini.
4. Selain itu, diharapkan hasil penelitian ini juga dapat dipakai sebagai referensi dibidang karya ilmiah, serta bahan masukan bagi penelitian dalam bidang yang sama di masa yang akan datang.

1.6 Sistematika Penulisan

Naskah ini disusun secara sistematis yang didasarkan pada pedoman penulisan tesis yang telah ditetapkan oleh Program Pascasarjana Universitas Atma Jaya Yogyakarta dengan urutan penyajian sebagai berikut :

BAB I PENDAHULUAN

Bagian ini berisi latar belakang, perumusan masalah, batasan masalah, keaslian penelitian, manfaat yang diharapkan, tujuan penelitian, dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Bagian ini memuat uraian sistematis tentang beberapa penelitian yang dihasilkan oleh peneliti terdahulu serta tinjauan pustaka yang relevan dengan penelitian ini.

BAB III LANDASAN TEORI

Bagian ini dapat berbentuk uraian kualitatif, model matematis, atau persamaan-persamaan yang langsung berkaitan dengan bidang ilmu yang diteliti.

BAB IV METODOLOGI PENELITIAN

Bagian ini berisi tentang uraian terinci mengenai bahan atau materi penelitian, alat dan langkah-langkah penelitian.

BAB V HASIL DAN PEMBAHASAN

Bagian ini memuat uraian secara jelas dan tepat mengenai hasil yang didapat dalam penelitian ini. Pembahasan berisi tentang analisis terhadap hasil yang diperoleh, dan kesulitan-kesulitan serta cara pemecahannya yang ditinjau secara utuh baik secara kualitatif, kuantitatif maupun normatif.

BAB VI KESIMPULAN DAN SARAN

Bagian ini menuliskan kesimpulan yang didapatkan dari penelitian ini serta saran untuk pengembangan ke depan.

